Probing the mechanisms by which endogenous mammalian transposons can disrupt gene expression

Honors Research Thesis

Presented in Partial Fulfillment of the Requirements for graduation
"with Honors Research Distinction" in the undergraduate
colleges of The Ohio State University

by
Tanvi V. Joshi

The Ohio State University
May 2014

Project Advisor: David E. Symer, M.D., Ph.D.
Human Cancer Genetics Program
Department of Molecular Virology, Immunology and Medical Genetics
Department of Internal Medicine

Table of Contents

Abstract

Around half of the mammalian genomes are composed of repetitive elements, of which most are transposable elements (TEs). Although certain TEs or "jumping genes" are still actively mobilized in the mammalian genome, their impact on gene expression remains largely unknown. We have focused on two genes, *Slc15a2* in mice and *MCC* in humans, both of which have been linked recently to TE-mediated gene disruption. This research seeks to study the mechanisms by which such endogenous TE polymorphisms can result in transcriptional disruption. In the case of *Slc15a2*, we previously reported that a polymorphic endogenous retrovirus (ERV) in intron 7 results in a 13-fold increase of a 1.2 kb prematurely-truncated transcript and a 39-fold decrease in the full-length, 4 kb transcript. This transcriptional disruption further resulted in a 3- to 9- fold down-regulation of the protein, PEPT2. We have also characterized antisense transcripts that were initiated from the bidirectional promoter at the 5' end of the ERV. In this study, we hypothesized that variable epigenetic regulation at the 5' long terminal repeat (LTR) of the ERV could be associated with differential antisense transcription, which in turn, could affect variable expression of *Slc15a2*. To demonstrate this, we used bisulfite sequencing to assess cytosine methylation levels at the ERV, demonstrating a correlation between decreased local DNA methylation and an increase in truncated transcripts at *Slc15a2*. In the case of human *MCC*, altered expression in liver cancer samples was recently attributed to three distinct intronic TE polymorphisms. However, our qRT-PCR analysis on a panel of 10 human B-lymphocyte lines from the Coriell repository did not validate this putative association between the presence or absence of TEs and varied expression of *MCC*. We have also screened 22 primary liver cancer samples for these TEs, to study potential variation in *MCC* expression, in tumor and matched-normal samples. Again, we could not confirm the reported correlation. We conclude that careful

case-by-case analysis is needed to evaluate the many possible biological impacts of TEs on gene expression. Our results suggest that in certain cases epigenetic controls may play a role in the mechanism of TE-mediated transcriptional disruption.

Lay Summary

Transposable elements are repetitive sequences that are capable of movement in mice and humans. Novel insertions can result in genetic instability by initiating disruption of normal gene expression and function. Several studies provide compelling examples of genes in which transposon integrations induce transcriptional and functional disruptions. In this research, we studied the previously reported transposon-mediated gene disruption at two genes, *Slc15a2* in mice and *MCC* in humans. We further probed some potential mechanisms by which these polymorphic transposable elements, shown to be present in the gene introns, can disrupt gene function. Transposons and their roles in gene disruption and potential disease formation remain largely unexamined. Therefore, conducting this study will provide insight into the mechanisms by which transposon-mediated gene disruptions may occur in mammals.

Introduction

Transposable elements (TEs), also known as "jumping genes" are mobile, repetitive elements that comprise around 50% of the human and mouse genomes[1, 2]. They are dispersed non-randomly across mammalian and other eukaryotic genomes[3]. Though most TEs are ancient relics that have become fixed in genomes over evolution, some are still actively mobilized and can initiate mutagenic effects[3]. In this research project, we have studied the biological impacts of TEs in mammals. We focused on two genes, *Slc15a2* in mice and *MCC* in humans, where strong TE-mediated impacts on transcription have been implicated recently[4, 5]. In particular, we sought to investigate the mechanisms by which polymorphic TEs can disrupt gene expression. Several well-characterized examples include *Agouti variable yellow (A^{vy}), CDK5 activator binding protein (Cabp)* and *Notch-1* in the mouse genome and provide compelling evidence to the robust biological impacts that TEs can exert on gene expression[6-8].

Endogenous TEs can be subdivided into two major categories: DNA transposons and RNA transposons[3, 4]. DNA transposons move by a "cut-and-paste" mechanism, where a transposase excises and reinserts the transposon in a new genomic location[3, 4]. RNA transposons, also termed retrotransposons, use a "copy-and-paste" mechanism[4]. They are transcribed and converted into cDNA by a reverse transcriptase and inserted into a new locus[3, 4]. Unlike DNA transposons, retrotransposons make up a much larger part of the mammalian genome and are still actively mobilized[3]. Therefore, they are the main topic of discussion here.

Classes of active, mammalian retrotransposons

The four classes of retrotransposons most prevalent in mammalian genomes are long interspersed elements (LINEs), short interspersed elements (SINEs), SVA elements and ERVs [3].

This thesis will study transcriptional disruptions mediated by the L1, *Alu*, and Intracisternal A Particle (IAP) elements, which are specific types of LINE, SINE and ERV elements, respectively. These three TE classes are relatively young and still actively mobile in the human and mouse genomes[1].

L1 is known as the most important, mobile TE in the human and mouse genomes, with over 500,000 copies dispersed throughout in each[9, 10]. They are capable of transcribing their own endonucleases and reverse transcriptases, enabling them to undergo autonomous transposition[9]. The majority of L1 insertions are 5' truncated and remnants of ancient retrotransposition events, which have become fixed in populations and accumulated many mutations[3, 9]. However, full-length human L1s contain a bidirectional promoter at the 5'untranslated region (UTR) and have been implicated in examples of gene disruptions[1, 3]. Recently, we characterized another antisense (AS) promoter in the ORF1 of full-length mouse L1 elements, which was shown to initiate many fusion transcripts within endogenous genes[13]. Furthermore, L1 promoter hypomethylation has been linked to disease phenotypes such as multiple myeloma, chronic lymphocytic leukemia, and chronic myeloid leukemia, suggesting that *de novo* L1 transcription and consequent transposition may be increased in sporadic human cancers[9].

SINEs have reached a copy number of over 1 million in the human genome[9]. Some young elements continue to be actively mobilized[9]. In the human genome, over 67% of SINEs contain *Alu* elements[12]. They are non-autonomous elements that must rely on other elements (such as LINEs) for retrotransposition[3]. *Alus* range from 100-400 bp in size and contain a strong internal promoter[1]. Most intronic *Alu* elements are enriched specifically near alternatively spliced exons and can thereby affect normal transcription in distinct ways[9, 12].

Another class of non-autonomous TEs is SVA elements, which are a composite of SINE-R, VNTR, and *Alu* sequences[14, 15]. Similar to *Alu* elements, they can be mobilized by L1 machinery and represent another class of TEs that is capable of active movement in the human genome[15].

ERVs are not mobile in humans, but result in at least 10% of germline mutations within the mouse genome[12]. Of these, the ERK-K family is a group of young, full-length TEs that have intact ORFs and are flanked by virtually identical long terminal repeats (LTRs), enabling them for active retrotransposition[4]. Particularly, Intracisternal A-Particle (IAP) family members are known to be mobilized autonomously. These elements are transcriptionally activated due to DNA hypomethylation[4]. Furthermore, we recently showed that chromosomal locations of novel ERV integrants are completely different in highly divergent mouse strains. For example, when we compared genomes of B6, CAST, and SPRET mice, we found only 4 out of the several thousands of IAPLTR1 integrants mapped being observed at orthologous loci[4]. This indicates that novel ERV integrants may be a driving factor in speciation and genetic diversity within the mouse genome.

Aberrant gene expression at various genetic loci such as *Rpo1-4, Slc15a2, MCC, Agouti, Cabp, Notch-1, and APC* in the mouse and human genomes has been attributed to novel TE integrations[4-7, 16]. Due to TEs' abundance and potential to induce variable gene expression, we contend that case-by-case analyses of such novel TE integrations and possible associations with altered transcript expression, structures, and functions will be fundamental to understanding their wide-ranging biological impacts.

TEs can introduce their own promoters and transcription factors and are thereby capable of initiating novel transcripts[3]. Recent findings indicated that between 6-30% of the entire mammalian transcriptome is initiated from TEs[17]. Indeed, around 18% of transcription start sites are found to overlap with repetitive elements such as L1s and SINES in the human genome[3]. For instance, in *DNA Methyltransferase-1 (Dnmt1)* mutant mice, 7 out of the 16 hypomethylation induced T-cell lymphomas were observed to contain a *de novo* IAP insertion in *Notch-1*[8]. This intronic integration drove the expression of a novel fusion transcript that induced the oncogenic function of *Notch-1*and contributed to lymphoma formation[8]. In addition, an HBV-human fusion transcript (HBx-LINE1) was found recently to be expressed in 23.3% of Hepatocellular Carcinoma (HCC) samples examined[18, 19]. The hybrid transcript was expressed as a long non-coding-RNA, which in an oncogenic environment was correlated to decreased patient survival and tumor formation in mice[18, 19].

Novel insertions of TEs also can initiate ectopic activation of genes as illustrated at the *Agouti viable yellow (A^{vy})* and *Axin-fused ($Axin^{Fu}$)* loci in mice[6, 20]. The integration of IAP elements within each of these two genes introduces a bidirectional promoter, which drives the ectopic activation of the two genes and leads to disease formation in mice[6]. The A^{vy} gene presents a compelling example of how promoter activity of a newly integrated TE can affect tissue-specific expression of a relatively innocuous gene, normally resulting in sub-apical yellow bands on black hair, into a disease-causing gene, resulting in its over expression in a range of tissues and causing a yellow coat color, obesity, and type II diabetes[21].

Endogenous TEs have also been known to introduce alternative splice sites by providing ectopic splice acceptor and donor sites[3]. Intronic ERVs can trigger alternative RNA splicing by introducing endogenous, cryptic polyadenylation signals, resulting in differential transcriptional

expression[3]. Such an effect was recently shown at two genes, *CDK5 activator binding protein (Cabp)* and *Solute carrier family 15, member 2* (*Slc15a2)*[4, 7]. In both cases, an intronic ERV initiates the premature termination of transcription by inducing the use of weak endogenous poly (A) signals that are typically not used in the normal expression of each gene[4, 7].

To investigate the mechanisms by which TEs can initiate the numerous examples of gene disruptions observed thus far, it is important to consider the factors known to affect gene expression more broadly in the mammalian genome. Typically, regulation of gene expression occurs primarily at the level of transcriptional and post-transcriptional gene regulation[22]. Transcriptional regulatory factors include DNA methylation, histone modifications, DNA-binding transcription factors and other chromatin-associated factors[23]. Post-transcriptional gene regulators include transcriptional interference, double-stranded RNA interference, alternative splicing, and other mRNA stability factors[24]. This research project will focus how DNA methylation and antisense transcription can contribute to transcriptional and post-transcriptional gene silencing, respectively.

Long non-coding RNAs (ncRNAs) including *cis*-natural antisense transcripts (*cis*-NATs) have recently been identified, forming a substantial portion of the mammalian transcriptome[25]. Recent reports indicate antisense transcription is widespread in cells and has been observed at over 50-70% of genetic loci[26]. While they have formerly been dismissed as "transcriptional noise", there is growing evidence that *cis*-NATs may play a regulatory role on gene expression corresponding to the complexity of organisms – from prokaryotes to eukaryotes[25, 27].

A majority of the observed *cis*-NATs are linked to alternative splicing events observed in sense-antisense (SAS) gene pairs[28]. In addition, albeit less frequently than the SAS pairs,

antisense transcripts have been reported to be induced by polymorphic TE insertions and lead to variable sense gene expression[29]. Recently, 13 cases of human-specific antisense transcripts initiated by ectopic promoters introduced by TE insertions have been documented. This study further suggested that the variation in sense gene transcription as a result of the observed antisense transcription contributed to the evolutionary divergence between humans and chimpanzees[29].

Two potential mechanisms by which disruption of gene transcription can be mediated by antisense transcription include: (1) transcriptional interference, by which interactions of two RNA pol II complexes moving in opposing directions hinder transcription of one or both strands and (2) the formation of a double stranded RNA molecule, which could lead to post-transcriptional silencing via RNA interference (RNAi)[29].

Transcriptional interference (TI) is the suppression of one transcriptional activity, directly and *in cis* by a second transcriptional activity[30]. TI is usually asymmetric with the promoter of one transcript being more active than that of the second[30]. Over 20% of total human transcriptional start sites in protein coding loci have been reported to occur within TE sequences[31]. Because most *cis*- NATs originating from intronic TEs are in the antisense orientation relative to host gene transcription, the two RNA pol II transcription complexes arising from convergent promoters can collide and result in early transcriptional termination of one or both of the transcripts[30, 31]. A study of a polymorphic mouse IAP integrant at *Cabp* suggested that antisense transcripts arising from its 5'-LTR promoter could result in premature transcriptional termination observed just upstream of the TE[7].

Antisense transcription can also mediate gene expression through the formation of a double-stranded RNA molecule (dsRNA), which could lead to RNAi[29]. RNAi occurs in many diverse eukaryotic organisms[32]. The dsRNA is cleaved into short fragments of about 21-nt by an RNAase-III-type-endonuclease, Dicer, creating what are referred to as short interfering RNAs (siRNAs) [32]. These complexes are then assembled into RNA-induced silencing complexes (RISCs), by which degradation of mRNA occurs and host gene expression is silenced[32]. RNAi remains a complicated and relatively unexplored component of post-transcriptional gene silencing[32].

In mammalian genomes, the primary role of DNA methylation has been to defend the mobilization and expression of TEs via transcriptional gene silencing[33, 34]. For instance, a high rate of TE-mediated gene disruptions was observed in the *D. Melanogaster* genome, which contains no genome defense in the form of DNA methylation[33]. In mammalian genomes, DNA methylation is controlled by a family of DNA methyltransferases that include *Dnmt3a*, *Dnmt3b*, and *Dnmt1* and establish methylation patterns at differing points during embryonic development[35]. These enzymes are involved in two distinct DNA methyltransferase (MTase) activities observed in all mammalian cells: 1) a maintenance MTase activity in proliferating cells, to ensure the maintenance of methylation patterns during cell division and 2) a *de novo* MTase activity to reestablish the DNA methylation patterns during the post-implantation stage[36].

Through embryonic development, there is a complete erasure of epigenetic marks in primordial germ cells[37]. In mouse embryos, around 12.5 days post coitum (d.p.c), *de novo* DNA methylation patterns are reestablished by *Dnmt3a* and *Dnmt3b* [35, 37]. Furthermore, tissue specific DNA methylation patterns are established during development by *Dnmt3a* and *Dnmt3b* and transmitted during cellular replication by the maintenance methylation activity, mediated by

Dnmt1[36]. We utilized genetic models of *Dnmt1* mutations to analyze correlations in epigenetic regulation and TE-mediated gene disruption.

As the maintenance MTase, *Dnmt1* has the highest expression level in vivo and mutations in this gene seem to produce the most severe phenotypes relative to the other family of DNA MTases[38]. *Dnmt1* is the best studied methylation enzyme in mammals and currently, four different mutant alleles have been generated: $Dnmt^{n}$, $Dnmt^{s}$, $Dnmt^{c}$, and $Dnmt^{chip}$ [39, 40]. This thesis makes use of $Dnmt1^{c}$ mutation, which is a true null allele for the gene and $Dnmt^{chip}$ which encodes a hypomorphic allele[40].

The $Dnmt^{c}$ mutation represents a deletion of two highly conserved motifs in the catalytic domain and causes complete inactivation of the MTase function[40]. No *Dnmt1* transcripts were detected in homozygous mutant mouse embryonic stem (mES) cells and no protein was observed[40]. The growth of homozygous mutant embryos ($Dnmt^{c/c}$) was arrested before the 8-somite stage and at 9.5 d.p.c homozygotes displayed a distorted neural tube and died during gestation[37, 40].

A rescue of the null $Dnmt^{c}$ allele by the expression of the gene's cDNA in its cognate locus was also reported[40]. The cDNA was inserted upstream of the mutation using the cDNA homologous insertion protocol, resulting in the $Dnmt1^{chip}$ allele. It is expressed from the gene's endogenous promoter, resulting in hypomorphic expression[40]. $Dnmt1^{chip/c}$ rescued mES cells still only express 10% of the DNA MTase relative to wild-type cells[41]. Further, compound heterozygous mice are runted, exhibiting only 70% of the body weight of WT mice, again corroborating the hypomorphic expression of the $Dnmt1^{chip}$ allele[41].

To gain insight into the underlying mechanisms of TE-mediated biological impacts within the mammalian genomes, we investigated correlations among DNA methylation, *cis*-NATs, and transcriptional disruption at *Slc15a2* and attempted to study these parameters at TE insertions in *MCC.*

Slc15a2 encodes for PEPT2, which is one of the integral membrane proteins that mediate cellular transport of oligopeptides and peptide-like drugs[42]. These proton-coupled transporters of the SLC15 family are phylogenetically conserved membrane proteins[42]. PEPT2 is highly expressed in the mouse kidney and brains[4, 42]. Though *Slc15a2* knock-out mice are viable and fertile, their physiological transport of certain oligopeptides such as beta-lactam antibiotics in the kidney and brain is significantly disrupted[42].

Our lab recently reported strong impacts by an intronic polymorphic ERV in disrupting expression at *Slc15a2*[4]. The presence of the ERV is strongly associated with 39-fold reductions in the 4-kb full-length transcript and a concomitant increase of 13-fold in a 1.2-kb truncated transcript (Figure 1A)[4]. This TE-mediated transcriptional disruption further results in a 3- to 9-fold down regulation of PEPT2 protein (Figure 1B)[4].

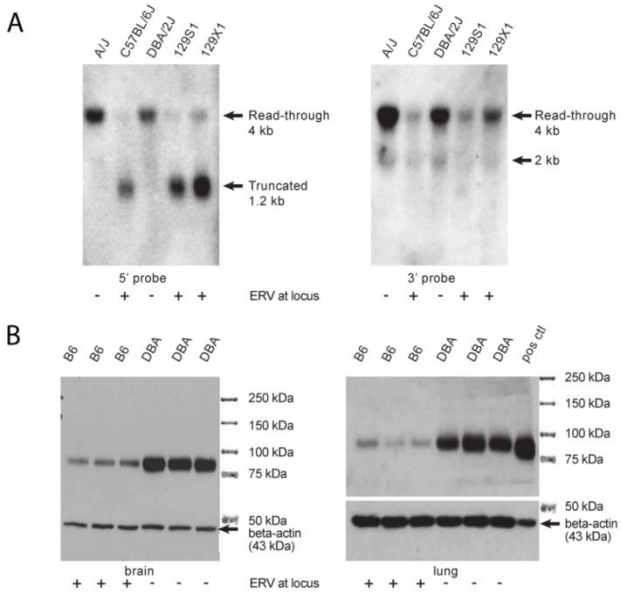


Figure 1: Polymorphic intronic ERV at *Slc15a2* triggers gene disruption. (A) Northern blots. ERV at locus results in expression of a truncated, 1.2-kb transcript. Variable transcription is illustrated using 5' probe (*left*) and 3'probe (*right*). No downstream 2-kb fusion transcript detected (*right*). (B) Western blots; the polymorphic ERV disrupts expression of the PEPT2 protein, encoded by *Slc15a2*, in brain (*left*) and lung (*right*). This figure is copied from Li *et al. Genome Research* (2012).

The polymorphic ERV in intron 7 of *Slc15a2* results in prematurely truncated transcripts terminating 1.5-kb upstream of the TE[4]. 5'-RACE cloning illustrated that the truncated transcript is terminated with a poly (A) tail sequence and contains part of intron 7 (Figure 2)[4]. Furthermore, this shorter isoform of the transcript does not contain any sequences of the ERV and terminates at two weak cryptic poly(A) signals[4].

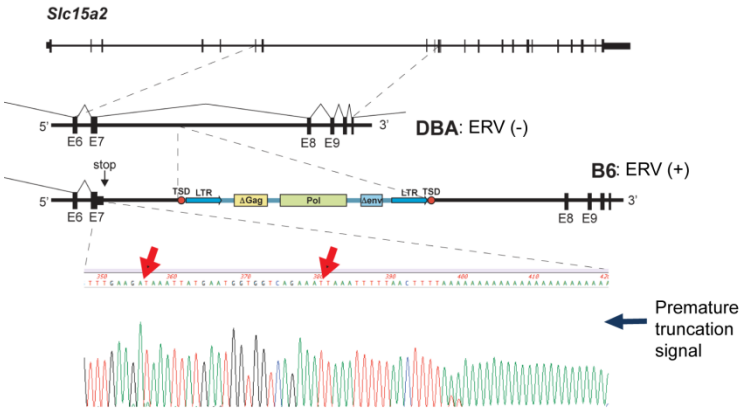


Figure 2: Gene schematic of *Slc15a2* in DBA and B6 mice. Polymorphic ERV present in B6 mice results in premature termination at intron 7 *(black arrow)*. Truncated transcript terminates at two cryptic poly(A) sites *(red arrows).* This figure is copied from Li *et al. Genome Research* (2012).

Another possible candidate gene where such gene disruption seems to be initiated by different polymorphic TEs is *MCC* [5]. Three novel TE integrations were correlated with strong inhibition of *MCC* in liver cancer patients[5]. *MCC* expression was shown to dramatically inhibit the oncogenic β-catenin/Wnt signaling pathway that is activated in hepatocellular carcinomas (HCC) [5, 43]; therefore, the TE-mediated down-regulation of *MCC* was indicated to result in the pathogenesis of HCC[5]. Four different cancer samples were either homozygous or heterozygous for the three TE polymorphisms in introns of *MCC* (Figure 3) [5].

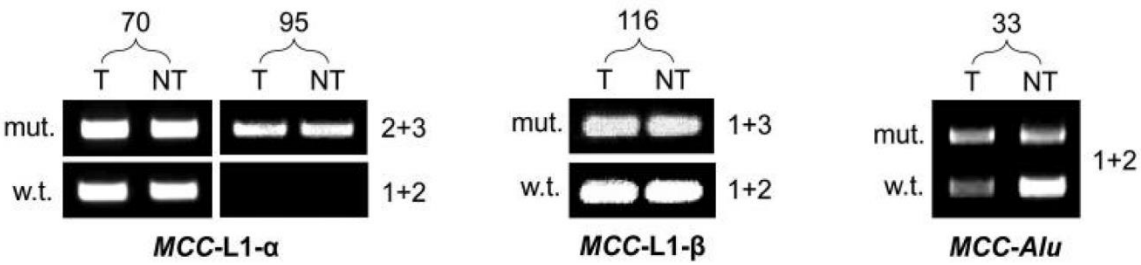


Figure 3: Site-specific PCR illustrates the presence of each TE-polymorph in the introns of *MCC* in tumor (T) and nontumor (NT) liver cancer samples. The presence of wt products indicates that sample is heterozygous for that TE. This figure was copied from Shukla *et al. Cell* (2013).

Two L1-Ta elements, belonging to the Ta subfamily of human-specific TEs known to actively retotranspose[9], were mapped in introns 2 and 7 and were termed L1-beta and L1-alpha, respectively (Figure 4) [5]. L1-beta is a full-length, 6-kb insertion, oriented antisense in relation to the gene. L1-alpha is truncated at 5.3-kb and is in the sense orientation[5]. Additionally, a novel *Alu* integration was observed in intron 4 in the antisense orientation in reference to *MCC* (Figure 4) [5]. The published results suggested that these three TE polymorphisms were the primary determinants of transcriptional disruption at the *MCC* locus in humans, although no mechanism was provided[5].

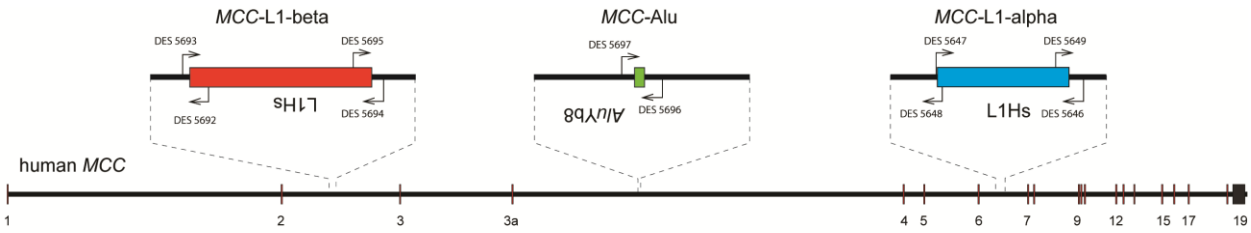


Figure 4: Gene schematic shows intronic TE polymorphisms. In humans, either one or none of the three intronic TEs depicted above is present at *MCC*. The L1-beta *(red)* and *Alu (green)* elements are full-length and in the antisense orientation relative to the gene. The L1-alpha *(blue)* is truncated and in the sense orientation. DES numbers indicate primers used to genotype the presence of each element in the TE-junction assay.

In both cases at *Slc15a2* and *MCC*, little is known about the mechanisms of the TE-mediated gene disruption that has been reported. We hypothesized that a correlation between DNA hypomethylation at the ERV$_{Slc15a2}$ would lead to an increase in truncated transcripts and decrease in read-through transcripts at *Slc15a2*. We proposed a model in which this local hypomethylation at the 5'-LTR of the ERV$_{Slc15a2}$ would initiate an increase in antisense transcription from the ERV promoter, leading to an increase in truncated transcript levels. As we describe below, we found that variable expression levels of the prematurely truncated transcript at *Slc15a2* were correlated to a decrease in DNA methylation levels at the ERV$_{Slc15a2}$. We further

determined that differential DNA methylation was associated with variable antisense transcription from the bidirectional promoter of the intronic ERV. Our results suggested that antisense transcripts initiated by the TE integration may lead to a post-transcriptional regulation of *Slc15a2* expression. By contrast, in the case of *MCC*, our preliminary data did not support an association between one of the three TE polymorphisms and the previously reported down-regulation of *MCC* expression.

Materials and Methods

Embryo collection

Two different procedures were used to set up mating cages between *Dnmt1*$^{c/+}$ and *Dnmt1*$^{chip/+}$ on the 129S1/SvlmJ and C57B1/6J strains.

(1) Procedure 1: Mating cages were set up on the first day of each week. Females were weighed and checked for vaginal plugs each morning before 10:30am. If plug was observed, the day was labeled as 0 day post coitum (d.p.c). Female and male mice were then separated and the female was weighed every day to confirm pregnancy. If plug was not seen at the end of the week, female and male mice were still separated and female was continuously weighed in case of pregnancy for 21 d.p.c.

(2) Procedure 2: Mating pairs were set up and maintained until a plug was observed. Each female mouse was checked for plugs and weighed each day. Weight gain and palpation were used to determine pregnancy.

Tissue Dissection and preservation

Embryonic tissues were collected using a dissecting microscope. Most tissues were divided in half to be used for RNA and DNA extractions. Brain, lung, and liver were collected for RNA. Tissues were kept in RNAlater (Ambion) at 4°C for 24h and were then frozen at -80°C for later use. Brain, lung, liver, and kidneys were collected for DNA. Tissues were directly kept in dry ice and frozen in the -80°C freezer. Embryonic limbs and tail were obtained for the purpose of genotyping.

Genotyping: DNA preparation and PCR

Sample Preparation: Tail and limb DNAs were extracted using our standard DNA isolation protocol. Samples were digested using Proteinase K and incubated at 55°C overnight. DNA was precipitated using 100% isopropyl alcohol and resuspended in 1x TE.

PCR: Three distinct conventional PCR assays were run to genotype *Dnmt1* wildtype (with primes DES3517 and DES3518), $Dnmt1^{c}$ (primers DES3420 and DES4821) and $Dnmt1^{chip}$ (DES3518 and DES3519) alleles. Conditions for the wildtype and $Dnmt1^{chip}$ assays include: Platinum Taq hot start enzyme and annealing temperature of 61°C and extension time of 80 sec. Conditions of the $Dnmt1^{c}$ PCR assay include: Platinum Taq hot start enzyme and annealing temperatures of 58 °C and extension time of 60 sec. (all primer sequences are detailed below)

RNA extraction

RNA was extracted using the TRIzol reagent protocol (Life Technologies). Samples were dried from RNAlater and homogenized using 2 mL of TRIzol reagent. 0.2 mL of chloroform was added for phase separation to 1mL of the homogenized tissue-TRIzol mixture. Samples were allowed to incubate for 2-3 minutes at room temperature. Tubes were centrifuged at 12,000 x *g* for 45 minutes at 4°C to allow the mixture to separate into a lower red phenol-chloroform phase, a white interphase, and a colorless upper aqueous layer containing the RNA. The upper, aqueous layer was carefully pipetted out and 0.6 mL of chilled 100% isopropanol was added. Samples were kept at -80°C overnight for RNA precipitation. Tubes were centrifuged for 30 minutes at 4°C at 12,000 x *g*. Supernatant was discarded to isolate the white pellet at the bottom. RNA pellet was washed twice with 1 mL of chilled 70% ethanol and centrifuged at 15,000 x *g* for 15 minutes each at 4°C. RNA pellet was air-dried for 5 minutes and resuspended in 25-30 µL of Nuclease-Free Water (Ambion), depending on pellet size. RNA purity was assessed by two

methods: 1) by using NanoDrop to quantify concentration and 260/280 and 260/230 ratios and 2) by running approx. 200 ng on a 1% agarose gel.

DNA extraction

DNA was extracted for bisulfite sequencing using the DNeasy Tissue Kit (Invitrogen) and the protocol provided was followed. This protocol is based on the lysis of cells and silica-based purification of DNA. Purity was assessed using NanoDrop.

qRT-PCR

Total RNAs extracted were synthesized into first-strand cDNAs for quantitative reverse transcriptase-mediated polymerase chain reaction (qRT-PCR) assays. 2.5 µg of total RNAs for each sample were primed for reverse transcription using oligo(dT) and SuperScript II Reverse Transcriptase (Invitrogen). qRT-PCR was performed using 1:20 dilutions of these cDNAs and Power SYBR Green PCR master mix (ABI) on a StepOnePlus instrument (ABI). Data were analyzed using the comparative CT method with an internal control house-keeping gene such as HPRT to normalize differences in sample loading.

For the *Slc15a2* read-through transcripts, primers DES4532 and DES4533 were used and truncated transcripts were quantified with DES4532 and DES4354.

To quantify *MCC* expression as reported (4), we used primers DES5933 and DES5934. For downstream *MCC* expression, primers DES5917 and DES 5918 were used.

Bisulfite Sequencing

Extracted embryonic brain DNA was bisulfite treated using the CpGenome Turbo Bisulfite Modification Kit (Millipore). In brief, 0.5 µg of genomic DNA was used for bisulfite modification; this involves the chemical conversion of unmethylated cytosine to uracil. DNA

was amplified using primers DES2650 and DES2651, which are specific for the 5'-LTR of $ERV_{Slc15a2}$. Amplicons were ligated into backbone of the TOPO- TA plasmid (Invitrogen) and transfected into One Shot TOP10 chemically competent *E. coli* cells (Life Technologies). 8-10 individual bacterial clones per embryonic sample were sent for sequencing through GeneWiz.

Transposon Junction PCR-Assay

Primers were designed to detect the presence and absence of each transposon element on a filled-/empty-target site basis. The 5'-junctions of L1-Ta elements were amplified to genotype positive samples. For the L1-alpha, primers DES5950 & DES5951 were used to amplify the 5'-junction and DES5949 &DES5950 were used for the empty target site (ETS). The L1-beta 5'junction was assayed using DES5692 & DES5693 and primers DES5693 & DES5694 were used for the ETS. For the short *Alu*, primers, DES5696 & DES5697 were designed in the flanking regions and amplified the entire element if present. Platinum Taq hot-start enzyme was used for all PCR reactions. PCR conditions for the L1-alpha: annealing temperature (Ta) of 58°C and extension time of 30 sec; for the L1-Beta and *Alu* elements: Ta of 60°C and extension time of 30 sec.

F9 Drug Treatment

The mouse teratocarcinoma stem cell line (F9) was cultured in a complete growth medium according to the ATCC protocol. F9 cells were treated with 2 µM 5-aza-dC, 300 nM Trichostatin A (an inhibitor of histone deacetylase, HADC) or with a combination of the two drugs for 24 hours at 37°C, 5% CO_2. After treatment, total RNAs of treated and untreated F9 cells were isolated following the TRIZOL RNA isolation protocol. First-strand cDNA from RNAs was synthesized by Superscript II reverse transcriptase (Invitrogen). RT-PCR was conducted using different PCR primer combinations to check the expression of *Slc15a2* read-through, terminated,

antisense from ERV 5' LTR and fusion transcripts from ERV 3' LTR. The expression of mouse *HPRT1* gene was a loading control. This work was conducted by Dr. Jingfeng Li.

Slc15a2 Minigene Constructs

Slc15a2 minigene construct was made by homologous recombination using the modified DH10B cell strain DY380. A/J and B6 mouse DNAs were amplified using PCR and co-transformed with the linearized plasmid into cells. To replace *Slc15a2* endogenous ERV with different promoters, plasmid pFN21A DNA and human and mouse genomic DNAs were amplified to get CMV, human L1.3 5'UTR, mouse L1 5'UTR, and mouse L1 antisense promoters. To create mutations at the *Slc15a2* cryptic poly (A) site, site mutagenesis was performed according to the Strategene Quick mutation kit. This work was conducted by Dr. Jingfeng Li.

RNA-Seq

We obtained RNA-Seq data for 14 mouse strains from Sanger Institute's mouse genome sequence project[43]. Unstranded RNA-Seq analysis was performed using 75 bp x 2 paired end reads with the Illumina GAII platform. We downloaded aligned RNA-Seq reads in BAM format (ftp://ftp-mouse.sanger.ac.uk) and compared expression levels of annotated genes and transcripts. We quantified the expression levels for genes annotated in the Ensembl database (v64). Normalization and comparison of samples were performed following the standard Cufflinks protocol[44].

Primer Name	Primer Sequence	Primer Description
DES3517	AGCCAGTTGTGTGACTTGGAAAC	*Dnmt1* wt genotyping
DES3518	CAATGATAGCTCTCTGGTGTGAC	*Dnmt1* wt & $Dnmt1^{chip}$ genotyping
DES3519	GCTTCCTTCTCAGCACCAG	$Dnmt1^{chip}$ genotyping
DES3420	AGGACCAAGGAAATGTGCTG	$Dnmt1^{c}$ forward (F) primer
DES4821	GCCTTCTATCGCCTTCTTG	$Dnmt1^{c}$ reverse (R) primer
DES4532	CTTGATTTCTATGTTCATCACACCC	*Slc15a2* exon 7 (F)
DES4533	ATCCTTTTCCACTCCACTCAC	*Slc15a2* exon 8, read-through transcript primer
DES4534	TGTACATCTTGCTTCCCATTGC	*Slc15a2* intron 7, truncated transcript primer
DES5672	TGAAGAGCTACTGTAATGATCAGTCAAC	*mHprt* (F)
DES5673	AGCAAGCTTGCAACCTTAACCA	*mHprt* (R)
DES2650	YGTAGTTTTGGTTTTGAAATGAAG	$ERV_{Slc15a2}$ 5'LTR, (F) bisulfite primer
DES2651	CATAAAAAAACTTCTAACAAACACTC	$ERV_{Slc15a2}$ 5'LTR, (R) bisulfite primer
DES5950	AGCTGCAGGTCTGTTGGAAT	*MCC* L1-alpha 5' junction
DES5951	CATGATGCTTCACCCCCTAA	*MCC* L1-alpha 5' flanking
DES5949	CACAGAGGTTACCCCTTGCTTTG	*MCC* L1-alpha 3' flanking
DES5692	AACTCCCTGACCCCTTGC	*MCC* L1-beta 5' junction
DES5693	GCATCATGAGTTCAGCCTCA	*MCC* L1-beta 5' flanking
DES5694	GGGGATTCTCTCTCTCGCTC	*MCC* L1-beta 3' flanking
DES5696	TCCCCACACCACTTACACAA	*MCC* Alu 3' flanking
DES5697	TTGAGTCATCATCCCAGTGC	*MCC* Alu 5' flanking
DES5933	TATGGAAACGACTCCTCGGC	*MCC* Exon 3a
DES5934	TCTCATGAGGTGGGACTGCT	*MCC* Exon 5
DES5917	CCCACTCACTTCAGGACTGC	*MCC* Exon 11
DES5918	TCCTCCAAGGTTATGGTCAGG	*MCC* Exon 12
DES4330	AACAGGGGACATAAAAGTAATTGG	*hHPRT* (F)
DES4331	GCGACCTTGACCATCTTTG	*hHPRT* (R)

Results

TE-mediated transcriptional disruption was recently shown to occur at two candidate genes, *Slc15a2* in mice and *MCC* in humans[4, 5]. We investigated these novel TE insertions to gain insight into the mechanisms by which such disruption in gene expression may occur.

We performed bioinformatics analysis of RNA-Seq expression data at *Slc15a2*, which attests to importance of looking at particular transcript variants to be able to detect TE-mediate gene disruptions. This work was performed by our bioinformatics expert, Dr. Keiko Akagi. The figure reveals that no difference in level of expression was observed when the average exon expression of *Slc15a2* across 14 diverse mouse strains was analyzed (Figure 5A).

By contrast, there was a very significant difference when studying full-length transcripts and truncated transcripts at the host gene in B6 versus the other 13 strains displayed here (Figure 5B & 5C). B6 mice are the only strain in this data that harbor the ERV insertion at *Slc15a2* (here after referred to as $ERV_{Slc15a2}$) (Figure 2).

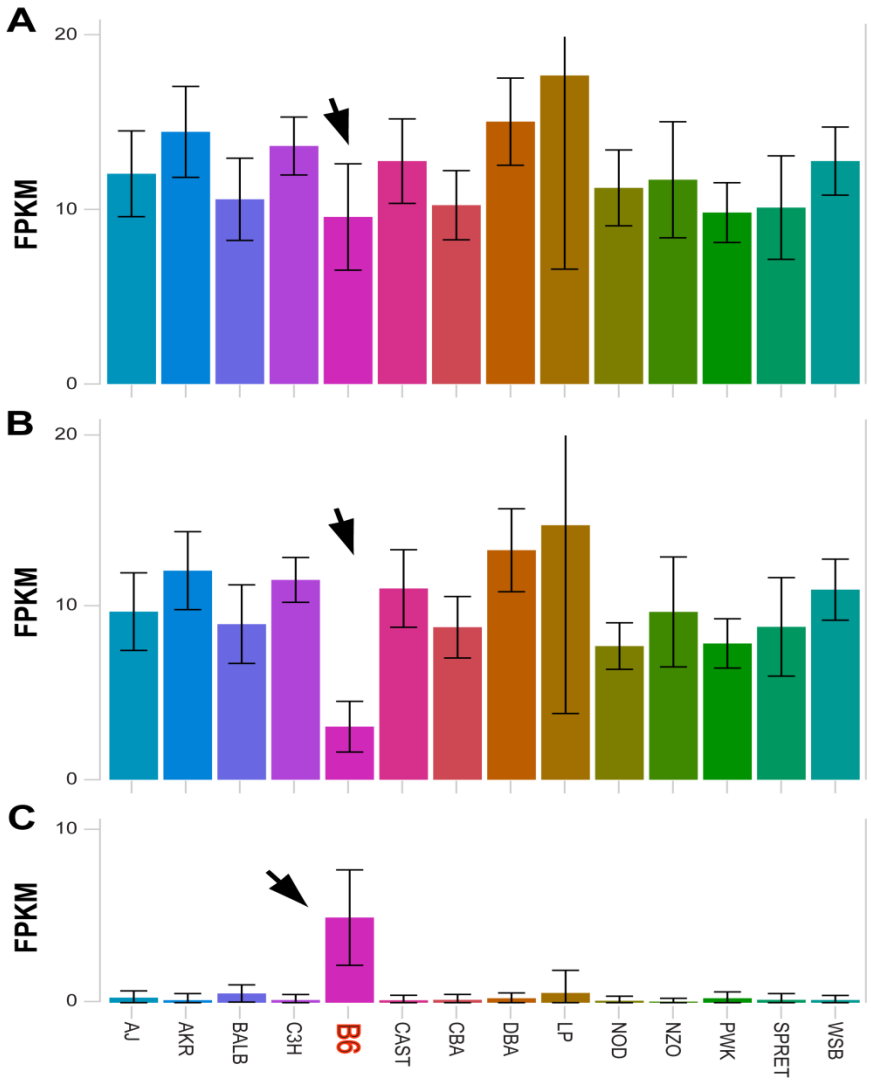


Figure 5: Transcriptional variation at *Slc15a2* in brain tissues of 14 diverse mouse strains, as determined from publicly available RNA-Seq data. (A) *Slc15a2* gene expression, by observing average expression across all exons in the host gene shows no variation in B6 *(arrow)* relative to the other strains. (B) Full-length transcript level expression analysis illustrates a significant decrease specifically in B6 mice (*arrow)*. (C) Expression of the *Slc15a2* truncated transcript was strongly increased in the B6 strain (arrow) due to the presence of the intronic ERV. *FPKM*, fragments per kilobase of exon per million fragments mapped.

An initial *in vitro* analysis was conducted in our lab by Dr. Jingfeng Li (unpublished data) to determine a possible role of epigenetic regulation at the $ERV_{Slc15a2}$ on *Slc15a2* expression levels. Mouse embryonal carcinoma (F9) cell lines typically do not express *Slc15a2* robustly (Figure 6). However, when cells were treated with 5-Aza-2-deoxycytidine (5-azadC), a drug that induces DNA demethylation, an increase in expression of both the upstream transcript and the prematurely terminated transcript at *Slc15a2* was observed (Figure 6). Results indicate that DNA methylation may play a central role in differential transcription mediated by $ERV_{Slc15a2}$.

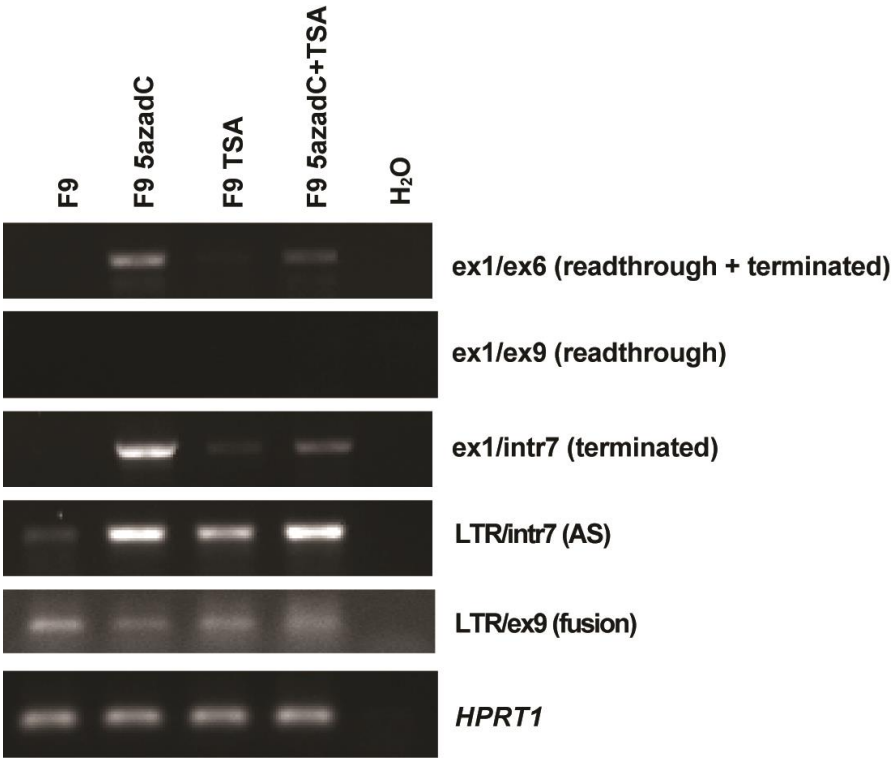


Figure 6: Increased expression of truncated transcripts in treated F9 cells, as visualized from RT-PCR assays of particular transcript structures (right). F9 cells do not express *Slc15a2* robustly as is illustrated by the absence of a band at both read-through and terminated transcripts. Upon treatment with the demethylating drug, 5-Aza-2-deoxycytidine (5-azadC), expression of the terminated transcript (ex1/intr7) was significantly increased. Concomitant increase in antisense transcription (LTR/intr7) from ERV promoter was also observed. F9 cells were also treated with Trichostatin A (TSA), a histone deacetylase inhibitor. However, antisense and truncated transcription was not as robustly increased as was the case with demethylation. This figure was generated by Dr. Jingfeng Li in the Symer laboratory.

To study whether this correlation occurs *in vivo*, we generated *Dnmt1*$^{chip/c}$ compound heterozygous mutants. Our previous assessments of DNA methylation levels in tissues of adult *Dnmt1* mutant mice revealed no significant decrease in methylation (data not shown). This is distinct from the initial characterization of *Dnmt1*$^{chip/c}$ mice, which described these mutants to be runted and to develop aggressive T-cell lymphomas due to severe hypomethylation[41].

We observed that *Dnmt1 chip* and *c* alleles were also not inherited at expected Mendelian ratios (Table 1). Thus, the lack of hypomethylation observed in adult *Dnmt1*$^{chip/c}$ mice and non-Mendelian inheritance of the two alleles indicates that they may be exerting a highly deleterious impact on embryonic development. Therefore, to obtain a hypomethylated mouse model for this study, we collected tissues from embryos at various developmental stages from intercrosses of *Dnmt1*$^{chip/+}$ and *Dnmt1*$^{c/+}$ parents. We expected an increased frequency in the *Dnmt1*$^{chip/c}$ compound-heterozygous mutants embryos relative to surviving adults (Table 1).

Intercross	*Mut/mut*	Mut/+	+/+ (wt)	N	P-Value	
Adult: *Dnmt1*$^{chip/+}$ *X Dnmt1*$^{c/+}$	3.7%	55.6%	40.7%	54	0	Non-Mendelian
Embryonic: *Dnmt1*$^{chip/+}$ *X Dnmt1*$^{c/+}$	17%	41.5%	41.5%	53	0.037	Non-Mendelian

Table 1: Non-mendelian inheritance of the *Dnmt1*$^{chip/c}$ genotype. Low allelic frequencies suggest that the two mutant alleles may be exerting a strong deleterious impact on embryonic development. This is further confirmed by the much lower allelic frequency observed in compound-heterozygous mutant (mut/mut) adult mice relative to that seen in embryos.

We conducted bisulfite sequencing to verify a hypomethylated state at $ERV_{Slc15a2}$ in *Dnmt1* mutant embryos. Bisulfite sequencing is a standard procedure that converts unmethylated cytosines in CpG dinucleotides to uracil and therefore, this method is used to determine the amount of methylated cytosine present[46]. We modified genomic DNA from various embryonic brain tissues of $Dnmt1^{chip/c}$ and wildtype embryos using bisulfite treatment. PCR products were Topo cloned (Life Technologies). We then amplified particular genomic regions using primers specific to the 5'-long terminal repeat (LTR) of the $ERV_{Slc15a2}$. Eight to 10 sequence reads were obtained per embryonic tissue sample. Our results indicated that $Dnmt1^{chip/c}$ embryos are on average 65% methylated relative to *Dnmt1* wildtype embryos, which exhibit an average of 95% methylation (Figure 7). Furthermore, larger differences in methylation were seen in wildtype versus mutant mice at younger developmental stages: 32.4% difference in methylated cytosines in 14 days d.p.c embryos in comparison to the 24.8% difference in day 1 newborn (NB) embryos (Figure 7). All embryos harvested were of either the 129S1 or B6 strain. Both of these mouse lineages are known to contain the ERV at *Slc15a2* in homozygosity[5].

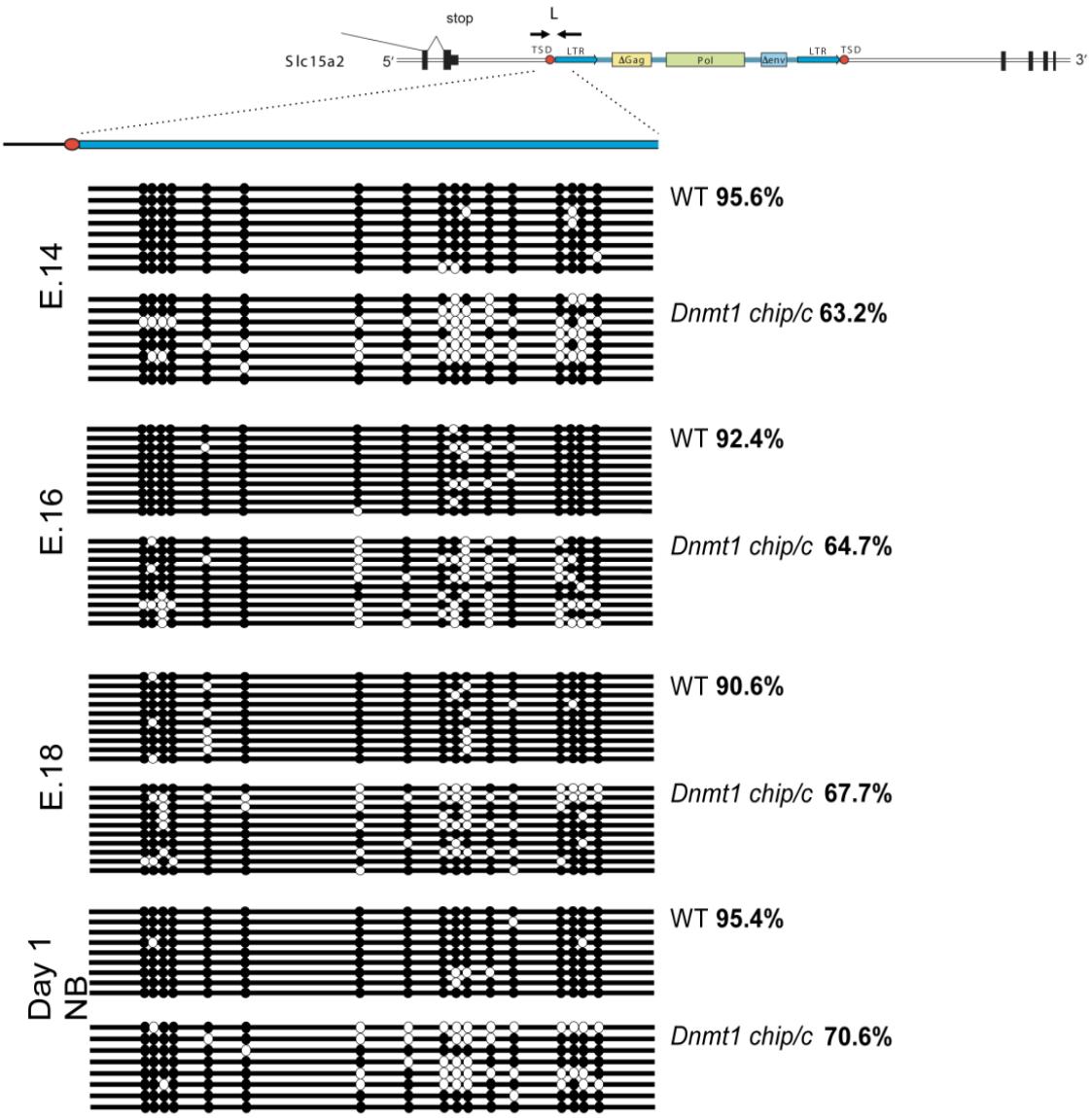


Figure 7: Validation of hypomethylation in *Dnmt1*$^{chip/c}$ mouse embryos relative to the wildtype genotypes. DNA methylation at the 5' LTR of the IAP was quantified using bisulfite sequencing of mouse embryonic brain DNA. Primers were used to PCR-amplify a 513-nt product, assessing 17 CpGs. Filled (black) circles and unfilled (white) circles indicate methylated and unmethylated cytosines, respectively. The number beside each panel represents the percentage of cytosines in CpG dinucleotides that were methylated in the amplicons measured.

Next, to study possible effects of altered epigenetic control at the $ERV_{Slc15a2}$ on variable expression of *Slc15a2 in vivo*, we compared RNAs isolated from wildtype and $Dnmt1^{chip/c}$ embryos, which we had validated to be hypomethylated through bisulfite sequencing. As noted above in bisulfite sequencing, we collected embryos at various developmental stages – ie. 14 d.p.c., 16 d.p.c, 18 d.p.c and day 1 NB.

Variable expression levels of the 1.2-kb truncated transcript and the 4-kb read-through transcript were quantified from RNA obtained from embryonic brain tissues using quantitative reverse transcriptase PCR (qRT-PCR). Relative expression of the read-through transcript was significantly down-regulated in hypomethylated embryos at all developmental stages investigated (Figure 8A). Conversely, we found that expression of the prematurely truncated transcript was up-regulated with a decrease in DNA methylation (Figure 8B).

This is consistent with our analysis in F9 cells, which illustrated that treatment with a demethylating drug leads to an increase in premature transcriptional truncation (Figure 1). However, expression of the truncated transcripts did not vary as significantly as expression of read-through transcripts. And as the difference in methylation levels between mutant and wt embryos decreased at later developmental stages, the variation in expression of truncated transcripts also decreased. For instance, in day 1 newborns, no significant difference in the expression of truncated transcripts is observed. In 14 d.p.c embryos, a greater difference in truncated transcript expression was observed between *Dnmt1* wildtype and mutant genotypes, which also exhibit the greatest difference in methylation levels. When we calculated the ratio of truncated transcripts to read-through transcripts, we noted a significant increase in $Dnmt1^{chip/c}$ embryos relative to wildtype embryos at all developmental stages (Figure 8C). This indicates that

the two transcriptional forms seen are inversely related to each other and both were associated with differential DNA methylation at the 5'LTR of ERV$_{Slc15a2}$.

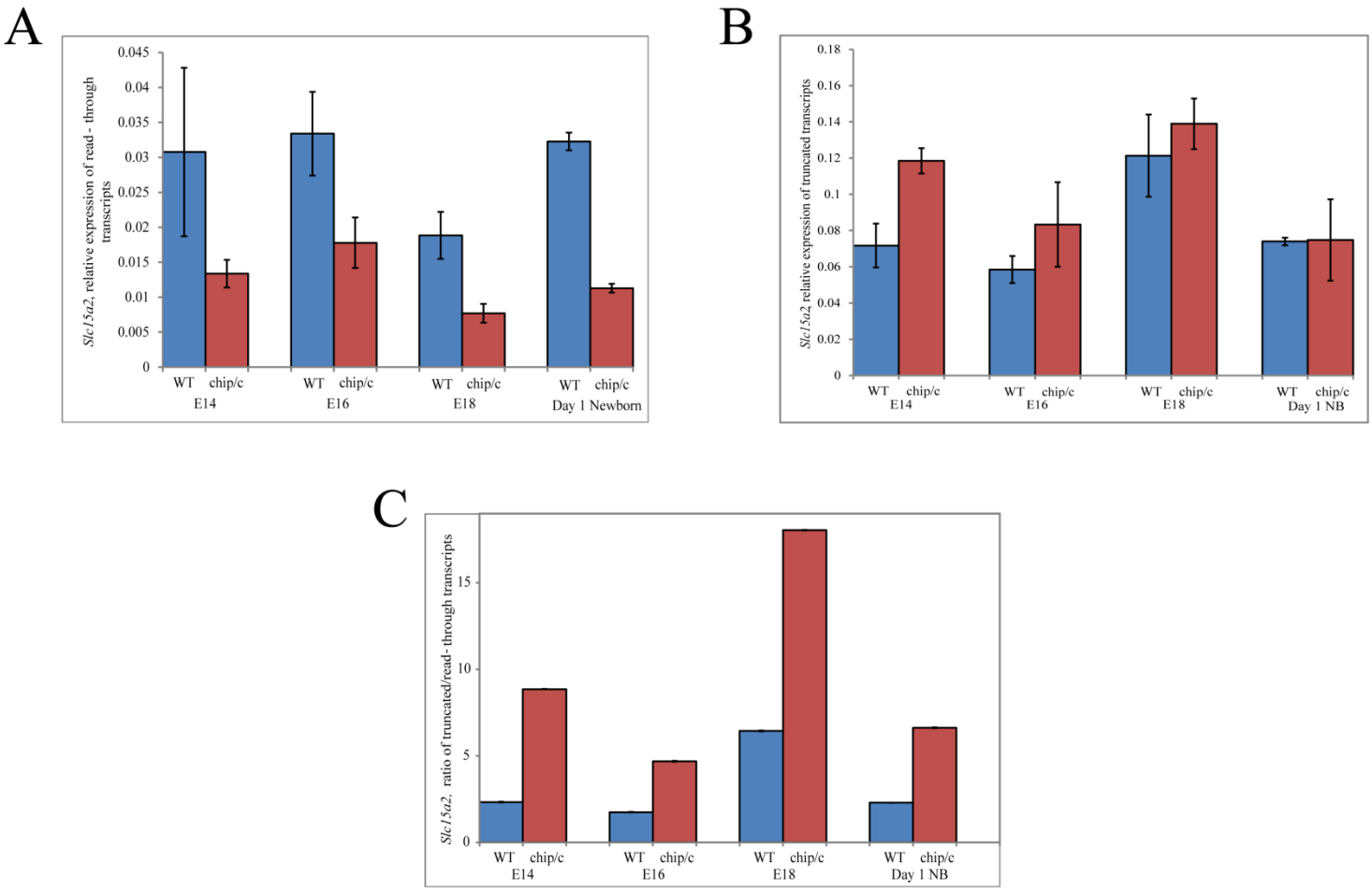


Figure 8: Expression of truncated transcripts is increased in hypomethylated embryos. qRT-PCR of extracted RNA from embryonic brain tissue of mice at different developmental stages. (A) Expression of read-through transcripts was significantly reduced in *Dnmt1* mutant embryos *(red)* relative to *Dnmt1* wt embryos *(blue).* (B) An increase in the amount of prematurely terminated transcript was observed with increasing DNA hypomethylation. However, truncated transcripts did not vary as significantly as read-through transcripts due to differential DNA methylation. No significant increase in truncated transcripts between $Dnmt1^{chip/c}$ embryos and wildtype in day 1 newborns (NB) was observed. (C) The ratio of truncated/RT transcripts significantly increased in hypomethylated embryos relative to wildtype embryos across all developmental stages.

The observed correlation of the increase in expression of truncated transcripts with DNA demethylation at the 5' LTR of $ERV_{Slc15a2}$ could result from antisense transcripts originating from the bidirectional promoter at the ERV. To investigate this model of transcriptional disruption, we studied changes in antisense transcript expression, DNA methylation, and truncated transcription in F9 cells. This work was part of the initial study conducted by Dr. Jingfeng Li from the Symer lab. His RT-PCR analysis showed that F9 cells treated with 5-azadC had increased expression of antisense transcription (unpublished data, Figure 6). This putative increase appeared to be correlated to an increase in terminated transcript expression, as mentioned previously (Figure 6). Meanwhile, untreated F9 cells with usual levels of methylation showed expression of neither the antisense transcripts nor the terminated transcripts (Figure 6). These data suggest that DNA methylation may be a major epigenetic regulatory factor that governs the disruption of transcription at *Slc15a2*, by potentially silencing antisense transcription from the polymorphic $ERV_{Slc15a2}$.

To study how increases in antisense transcript expression affect normal sense expression of the host gene, we generated a minigene construct, in which we manipulated the activity of the bidirectional promoter at the 5'LTR of the $ERV_{Slc15a2}$. This work was conducted by Dr. Jingfeng Li from our lab. The minigene construct contained a CMV promoter to replace the native gene's promoter, part of exon 1, exon 6, intron 6, exon 7, intron 7 and exon 8 of *Slc15a2* and a poly(A) signal (Figure 9A). The $ERV_{Slc15a2}$ was replaced by various different promoters: CMV, a mouse L1 5' UTR promoter, the mouse L1 antisense promoter, and the mouse L1 5' UTR promoter in the sense orientation to the gene (Figure 9A). Our qRT-PCR results analyzing the expression of prematurely truncated transcripts in relation to the strength of the antisense promoter indicated that in the introduction of a strong promoter such that from CMV, replacing $ERV_{Slc15a2}$ increased

antisense transcription and led to a significant increase in the prematurely truncated transcript at *Slc15a2* (Figure 9B).

Conversely, when a promoter was introduced in the sense orientation, we observed no expression of the terminated transcript (Figure 9B). This indicates that there is a strong correlation between antisense transcription at the $ERV_{Slc15a2}$ and premature transcriptional disruption at *Slc15a2*. It should be noted that in the presence of no $ERV_{Slc15a2}$, there is still a small amount of terminated transcription present (Figure 9B). This could be attributed to the presence of two weak cryptic poly(A) termination signals just downstream of exon 7, where the truncated transcript is terminated.

To assess the importance of the stronger one of the two cryptic poly(A) sites, we used site-specific mutagenesis to manipulate the same minigene construct (Figure 10A). In the presence of the endogenous poly(A) signal, there was some expression of terminated transcripts, even in the absence of $ERV_{Slc15a2}$ (Figure 10B). However, when the poly(A) signal was mutated, no terminated transcription was observed in the absence of the $ERV_{Slc15a2}$ and a reduced level of truncated transcripts is observed when the $ERV_{Slc15a2}$ was present (Figure 10B). Although it was reduced, the amount of truncated transcript still present may be attributed to the presence of the second weak poly(A) site.

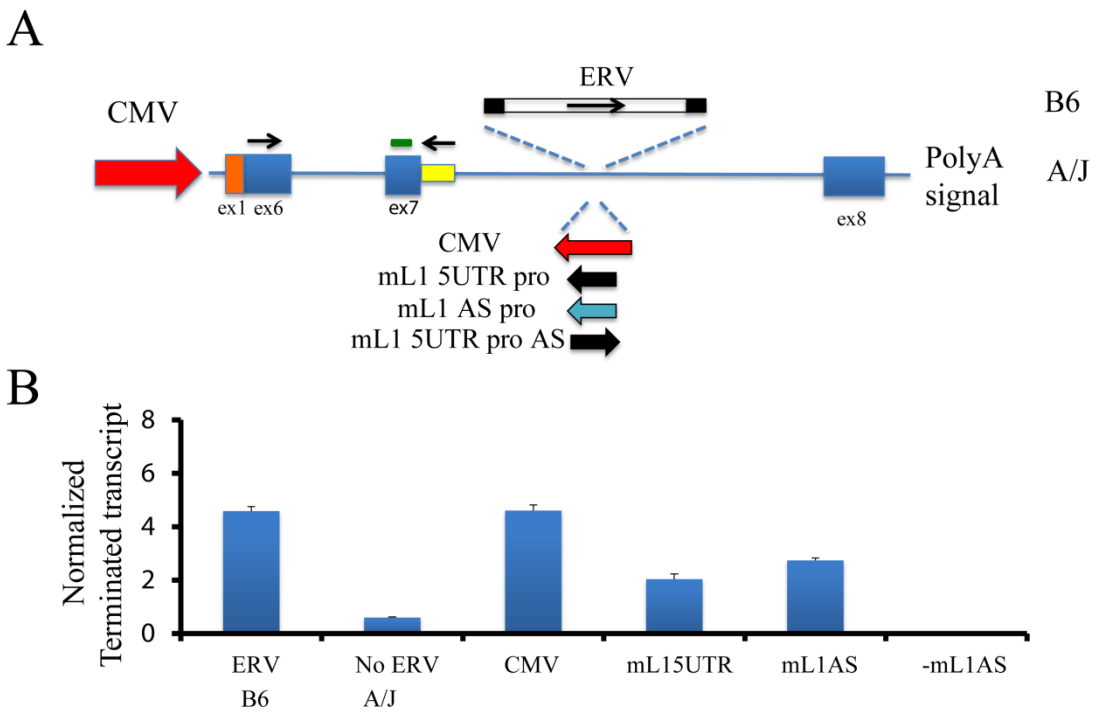


Figure 9: Antisense promoter activity is correlated to level of truncated transcript expression in our analysis of *Slc15a2* transcripts in a mini gene construct. The construct includes part of exon 1, exon 6, intron 6, exon 7, intron 7 and exon 8 sequences from the native *Slc15a2* gene in mice. The native gene promoter was replaced by a CMV promoter and a poly(A) termination signal was added to the end of the construct. (A) The $ERV_{Slc15a2}$ was replaced by various promoters of differing strength. (B) qRT-PCR of terminated transcripts at *Slc15a2*. In the presence of a strong promoter such as CMV, through which antisense transcription is increased, the expression level of truncated transcripts is significantly amplified. When the L1 5'UTR promoter is present in the sense orientation relative to the native host gene, no truncated transcripts are observed due to the absence of antisense transcription from the $ERV_{Slc15a2}$. This figure was generated by Dr. Jingfeng Li in the Symer laboratory.

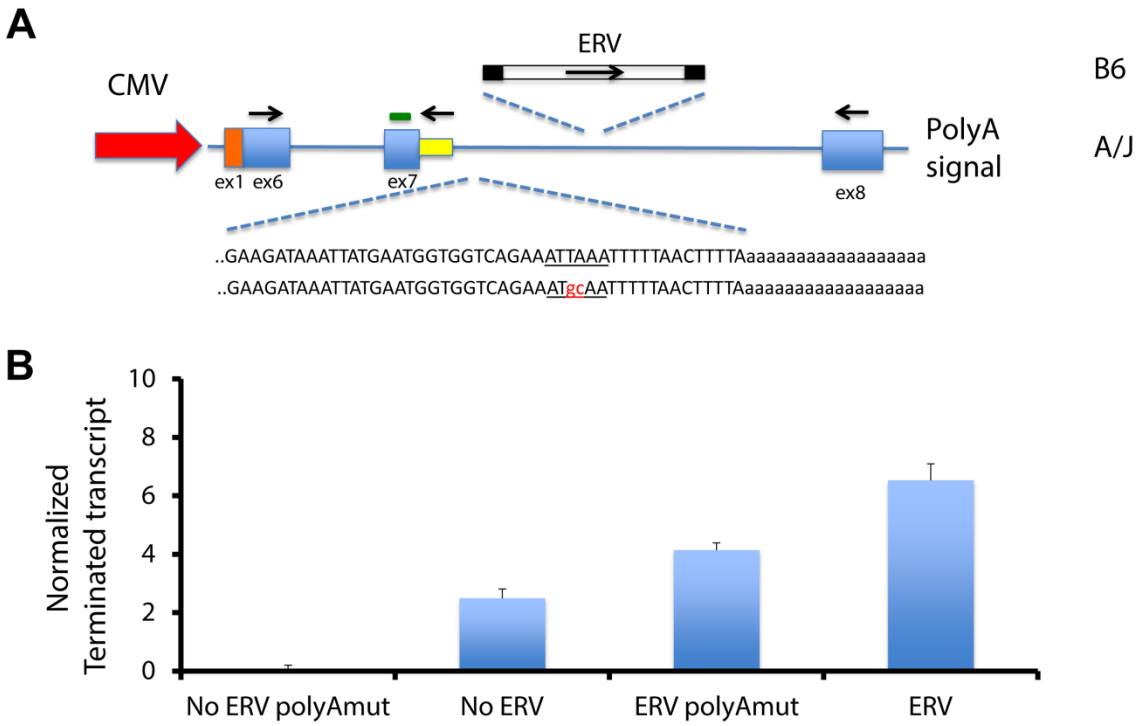


Figure 10: Cryptic poly(A) signals are important for transcriptional termination at *Slc15a2*. (A) Minigene construct (as described above) with poly (A) signal sequence mutated as shown *(indicated in red).* (B) qRT-PCR of terminated transcripts. In the presence of a mutation at the poly(A) site, reduced transcriptional truncation is observed. Some expression of this transcript was still observed in the presence of the ERV$_{Slc15a2}$ and poly(A) mutation, which could be attributed to the presence of a second, weaker poly(A) site upstream. This figure was generated by Dr. Jingfeng Li in the Symer laboratory.

Another example of TE-mediated gene disruption that was recently published is *Mutated in Colorectal Cancer (MCC)* in the human genome[5]. The presence of three polymorphic TEs in the introns of *MCC* was shown to contribute to significant down-regulation of the tumor suppressor in liver cancer patients[5]. Dr. Keiko Akagi, a bioinformatics expert in our lab, checked publicly available Complete Genomics datasets for the frequencies of the three TEs at *MCC*. We found that the allele frequency for each of the reported TE-polymorphisms was 9.1% for L1-beta, 1.9% for *AluYb8*, and 21.2% for the L1-alpha insertion. The relatively high frequency with which the three TE integrants are present at the *MCC* locus in humans – over a total of 30% - within the samples analyzed raised a question about whether the presence of each TE actually could lead to the putative strong down-regulation in gene expression. Therefore, our first goal was to validate whether this phenomenon actually occurs within humans diagnosed with liver cancer. If confirmed, we then would study the mechanisms by which this TE-mediated gene disruption occurred in this important tumor suppressor gene.

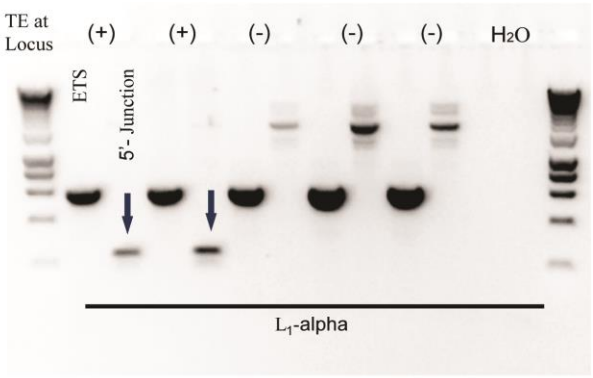
To first validate the presence of the three TEs at *MCC*, we obtained 10 human lymphoblastoid cell lines, previously known to contain one or none of the TEs and 22 human liver cancer samples – ie. tumor and matched normal sample. A TE-junction assay was developed for each TE, with transposon sequence-specific and degenerate primers for genomic PCR amplification. Two sets of primers were designed to amplify the 5'- and 3'- junctions of the L1-alpha element present in intron 7 of *MCC*. The 3'-junction primers failed to cleanly amplify positive samples and produced many non-specific products. Samples were deemed to be positive for this element upon the presence of approximately a ~230-bp product from the 5'-junction of the TE. A third set of primers was designed to amplify an empty target site (ETS) on alleles that contained no intronic TEs at *MCC*. All samples including those positive and negative for the L1-

alpha contained a 550-bp product, indicating the presence of an allele without TE integration. We validated 3/10 lymphoblastoid cell lines (data not shown) and 4/22 (18.2%) of human liver cancer samples to be heterozygous for the L1-alpha element (Figure 11A). All but one of the lymphoblastoid cell lines previously predicted to contain the L1-alpha were successfully validated with this junction PCR assay. We do not know the reason for this discrepancy.

A similar TE-junction assay was designed to determine the presence and absence of the L1-beta element in intron 2 of *MCC*. Again, the 3'-junction assay did not cleanly amplify the targeted sequence and produced many non-specific bands. All samples containing a 260-bp product from the 5'-junction of the TE were positive for the L1-beta. And all negative and positive samples produced a 350-bp product for the ETS assay. We confirmed the presence of the L1-Beta in heterozygosity in 2/10 lymphoblastoid cell lines and in 3/22 (13.6%) of liver cancer samples (Figure 11B).

For the *Alu*-Y element, two primers were designed flanking the element. Samples heterozygous for the *Alu* contained a 220-bp band from the allele negative for the TE and a 470-bp product indicating the presence of the element. 1/10 lymphoblastoid cell lines and 2/22 (9.1%) liver cancer samples were heterozygous for the *Alu* element in intron 4 of *MCC* (figure 11B).

A



B

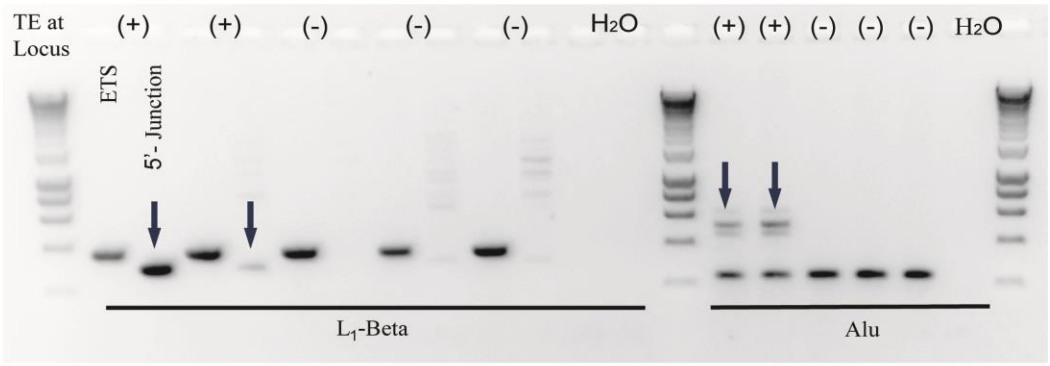


Figure 11: Validation of intronic TE polymorphisms at *MCC*. TE-junction PCR assay was conducted in liver cancer samples. (A) Positive elements for the L1-alpha contain a ~230-nt product *(indicated by blue arrows)* in the 5' junction assay. Each sample checked had a ~550-nt band for the ETS, indicating that the L1-alpha was present in heterozygosity. (B) Samples are positive for the L1-beta element if a ~260-nt product was present in the 5'Junction assay. All samples contained a ~350-nt band for the ETS, illustrating the presence of the L1-beta in heterozygosity. Samples that are heterozygous for the *Alu* element contained a ~470-nt for the positive-TE allele and a ~220-nt for the unfilled or TE-negative allele.

To validate the putative correlation between the presence of one or none of the three polymorphic TE and down-regulation of *MCC* expression, we synthesized the same qRT-PCR primers used previously[5]. The two primers were designed in exon 3a and exon 5 –downstream of the L1-Beta and upstream of the *Alu* and L1-alpha TEs[5]. We also obtained additional primers across all exons of *MCC*.

Our analysis did not confirm the correlation previously reported[5] between the presence of one of the intronic TE polymorphs and variable expression at *MCC* (Figure 12A). We also checked *MCC* expression levels at exon 11 and exon 12, downstream of all three TE polymorphisms. This assay illustrated similar results. Again, no consistent down-regulation of *MCC* due to the presence of one or none of the TEs was observed (Figure 12B).

Because TE-mediated *MCC* downregulation was previously examined in liver cancer patients[5] we next obtained 22 liver cancer samples to investigate this reported correlation. We conducted a similar qRT-PCR analysis on 8/22 samples that were previously validated to be positive or negative for each TE polymorphism, by using the same primers in exon 3a and exon 5 (as described above). Again, as was the case in lymphoblastoid cells, we observed no correlation between *MCC* expression and the presence of one or none of the three intronic TE polymorphs (Figure 13A & 13B).

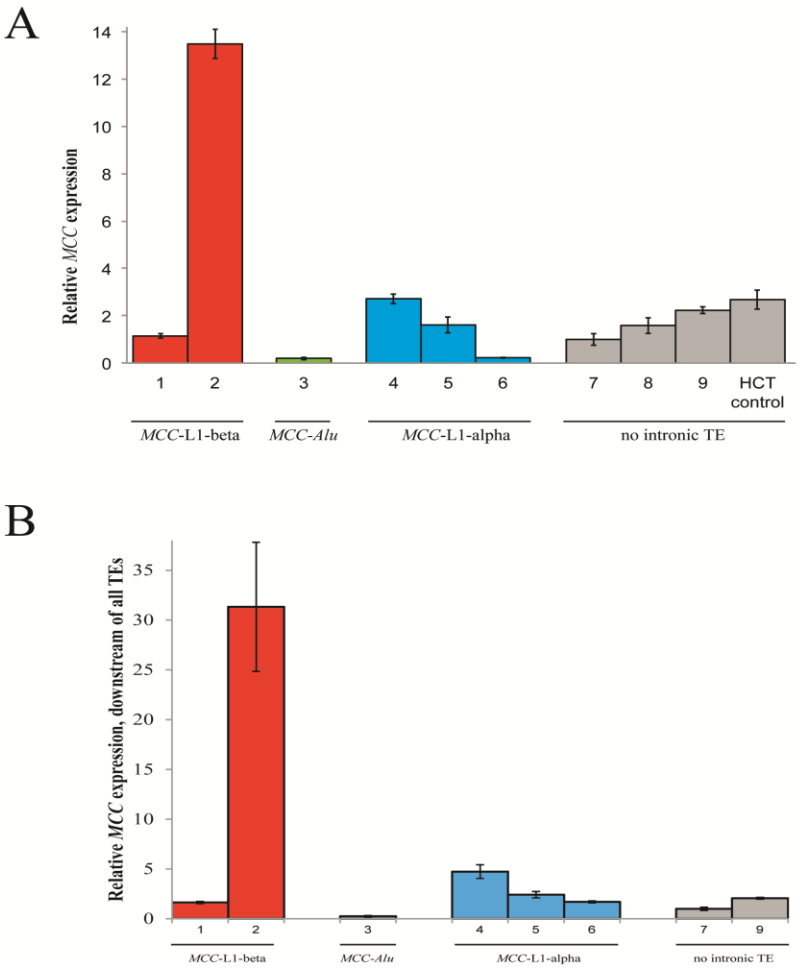


Figure 12: *MCC* expression does not correlate to the absence or presence of an intronic TE in our analyses of 10 human lymphoblastoid cells. (A) qRT-PCR analysis quantified expression at from exon 3a-exon5. (B) Expression quantified downstream of all intronic TE polymorphs in exons 11 and 12 of *MCC* illustrated results similar to those observed upstream in the gene.

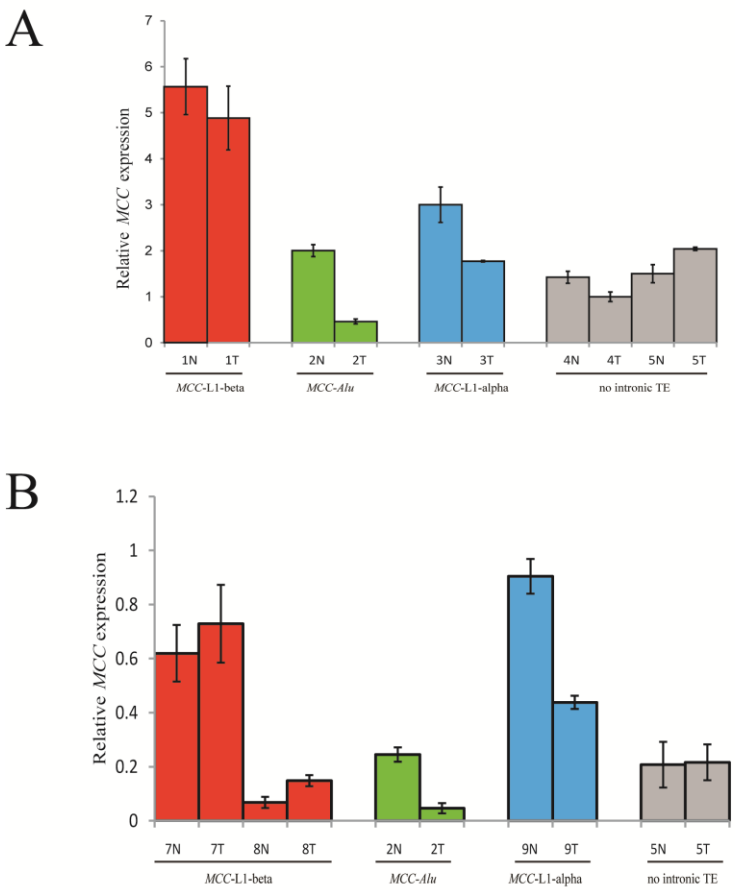


Figure 13: Relative *MCC* expression in liver cancer DNA – tumor (T) and matched-normal (N). Presence of polymorphic intronic TEs does not correlate to a down-regulation of *MCC.* (A) qRT-PCR analysis. 5 different liver cancer samples containing either one or none of the three TEs were analyzed for *MCC* expression. (B) Two additional liver cancer samples positive for L1-beta *(red)* and one additional for L1-alpha *(blue)* were analyzed.

Discussion

In this thesis, we evaluated the roles played by TEs in disrupting gene expression by studying two separate genes, *Slc15a2* in mice and *MCC* in humans. In the case of *Slc15a2*, our lab had previously defined the premature transcriptional termination that occurs as a result of an intronic polymorphic ERV[4]. This transcriptional variation was further shown to cause a disruption in gene and protein function[4]. A conceptually similar phenomenon, whereby intronic TEs could cause gene variation and functional disruption, was subsequently described at the *MCC* locus in humans[5]. We investigated both loci to gain an understanding of the mechanisms by which intronic transposons may be causing the reported transcriptional disruption.

RNA-Seq data of *Slc15a2* expression illustrated the importance of investigating transcript variants of genes at which TE-mediated transcriptional disruption has the potential to occur. An average expression level across all exons of *Slc15a2* did not show a difference across the 14 diverse mouse strains depicted (Figure 5). However, upon further analysis of full-length transcript and truncated transcript levels, a significant difference was observed in B6 mice in comparison to the other 13 strains (Figure 5). Our results point to the value of studying transcript variants in order to gain insight into the mechanisms by which gene disruptions can be initiated by intronic TEs.

DNA methylation is known to be highly enriched at CpG dinucleotides in TEs across the mammalian genome and typically acts as a form of genomic defense to inhibit TE-mediated gene disruption[25, 27]. We found this to be true in our study of the mechanism of ERV-mediated disruption of *Slc15a2* transcription. Our analysis in mouse embryonal carcinoma (F9) cells provided evidence that DNA methylation may act as a major form of regulation at the ERV$_{Slc15a2}$.

Upon treatment with a demethylating drug, 5-Aza-2-deoxycytidine, F9 cells exhibited a significant increase in expression of both the upstream sense transcripts and the prematurely truncated transcript at *Slc15a2* (Figure 6).

In studies of the classic *Agouti* gene in mice, it has been reported that *Dnmt1* is important in the maintenance of IAP methylation during embryonic mouse development[21]. We studied this effect *in vivo* by generating mouse embryos and newborn pups that were hypomorphic for *Dnmt1*. We sought to study expression of *Slc15a2* in *Dnmt1*$^{chip/c}$ mice, which were established beforehand to be hypomethylated[32]. However, our attempts to obtain a hypomethylated mouse model had not succeeded thus far. Unlike reports indicating that *Dnmt1*$^{chip/c}$ mice develop thymic tumors at about 4-8 months of age and generally do not survive past 10 months of age[32], our mutant mice were healthy well beyond one year. Bisulfite sequencing in these adult mice further illustrated that there was no significant DNA hypomethylation (data not shown). Our analysis also showed that the two mutant alleles, *Dnmt1 chip* and *c* were not inherited at expected Mendelian ratios (Table 1). Together, these findings indicated that the *Dnmt1*$^{chip/c}$ genotype may be posing deleterious impacts during embryonic development.

Therefore, we collected embryos at various developmental stages from *Dnmt1*$^{chip/+}$ and *Dnmt1*$^{c/+}$ intercrosses. As we anticipated, there was an increase in the frequency with which the *Dnmt1*$^{chip/c}$ genotype was observed when compared to surviving adult genotypes (Table 1). We validated hypomethylation by bisulfite sequencing, with the use of primers specific to the 5'LTR of the ERV$_{Slc15a2}$. *Dnmt1* mutant embryos displayed about 25-30% reduction in methylation at this LTR in comparison to *Dnmt1* wt embryos (Figure 7). Furthermore, we saw a slight reduction in DNA methylation levels in embryos at earlier developmental stages; mutant and wildtype embryos at 14 d.p.c were less methylated than day 1 newborn embryos. This is consistent with

recent reports which indicated that a wave of demethylation occurs just prior to implantation around stage 12.5 d.p.c.[37]. All embryos were bred on the 129S1 and B6 strains, which are known to be homozygous for the $ERV_{Slc15a2}$ [4]. Therefore, the observed variation in local hypomethylation at $ERV_{Slc15a2}$ in mutant embryos allowed us to investigate the effects of epigenetic regulation and differential transcription at *Slc15a2*.

Our qRT-PCR analysis studying the expression of truncated and read-through transcripts illustrated a significant down-regulation of read-through transcripts in hypomethylated embryos (Figure 8). A concomitant increase in prematurely truncated transcripts was observed with the decrease in DNA methylation (Figure 8). We acknowledge that more mutant embryos are necessary to corroborate our preliminary analysis of DNA methylation variation *in vivo*. However, our results thus far indicate that a reduction in methylation at the 5'LTR of the $ERV_{Slc15a2}$ was associated with the observed TE-mediated disruption of gene transcription. This provides us more insight about the mechanisms by which TEs can result in genomic mutagenesis. We confirmed that DNA methylation is important in the epigenetic silencing of TEs and that it acts as a form of genomic defense to prevent TE-mediated gene disruption in the mammalian genome[33].

Because TEs can introduce ectopic promoters, they are known to initiate their own transcripts – fusion transcripts or antisense transcripts that can disrupt host gene function[3]. The 5' LTR of the $ERV_{Slc15a2}$ contains a bidirectional promoter, which was previously shown to initiate transcripts antisense to *Slc15a2* [4]. Along with the increase in truncated transcripts that was seen in our initial analysis in demethylated F9 cells, a simultaneous increase in transcripts arising from the antisense promoter at the $ERV_{Slc15a2}$ was observed (Figure 6).

Next, we sought to manipulate the strength of this antisense ERV promoter and investigate the effect of variable expression of antisense transcripts. Therefore, we generated a minigene construct, containing essential parts of the native *Slc15a2* (Figure 9A). When expression of the antisense transcripts was increased by replacing the native ERV promoter with a strong, CMV promoter, the expression of the prematurely truncated transcript was augmented (Figure 9B). Conversely, when replaced by a promoter oriented in the sense direction to gene, whereby no antisense transcripts were generated, no truncated transcription was expressed (Figure 9B). Our data revealed that antisense transcription from $ERV_{Slc15a2}$ is positively correlated to the level of truncated transcription at *Slc15a2.* This suggested that transcriptional interference may be occurring at *Slc15a2,* in which two convergent promoters send two transcripts toward one another and result in the early termination of one of the transcripts[30]. This model is similar to one that was suggested to occur at the *Cabp* locus in mice[7]. However, this field of study remains largely unexplored.

Furthermore, we reported that the endogenous cryptic poly(A) signal present in intron 7 of *Slc15a2* is important in the presence of the $ERV_{Slc15a2}$ to initiate the increase in truncated transcript expression (Figure 10). Mutations at this poly(A) site resulted in a strong decrease in terminated transcript expression, indicating that the $ERV_{Slc15a2}$ induces the use of this weak termination signal to disrupt host gene expression. However, a small amount of truncation was still observed in the absence of $ERV_{Slc15a2}$ and this could be explained by the presence of another, weaker poly (A) site at that locus, as was previously described[4].

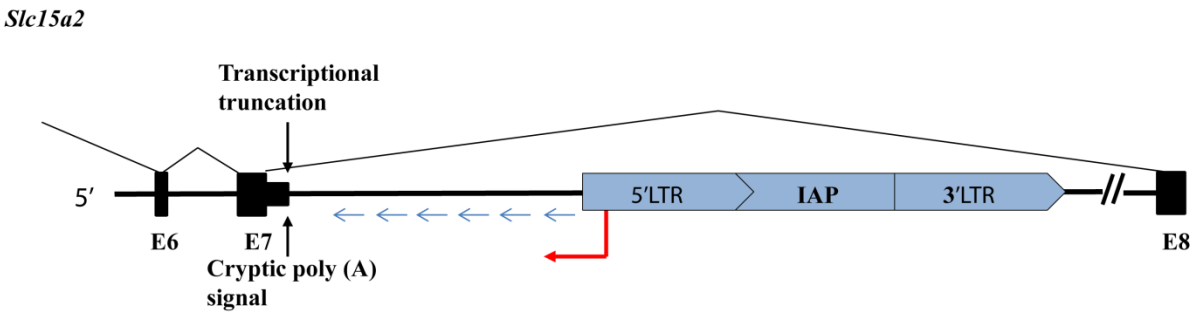


Figure 14: Proposed model of TE-mediated transcriptional disruption occurring at *Slc15a2*. Antisense promoter at the 5'LTR *(red arrow)* initiates antisense transcripts *(blue arrows)*. Transcriptional interference could induce premature truncation at *Slc15a2* at a cryptic poly(A) signal *(black arrow)*.

With this evidence, we propose a model for the mechanism by which the premature transcriptional truncation mediated by the ERV$_{Slc15a2}$ could occur under hypomethylated conditions (Figure 14). LTRs in the ERV in intron 7 of *Slc15a2* introduce an ectopic promoter within the native gene. Antisense transcripts initiated from this promoter could then result in the observed premature termination of *Slc15a2* transcripts in two main ways: (1) transcriptional interference, which occurs when the two RNA polymerase II complexes moving towards each other collide and (2) RNA interference, during which the formation of a double-stranded RNA molecule leads to post-transcriptional silencing of gene expression[29].

Further investigation will be required to gain additional insight into just how TE-initiated antisense transcripts may be mediating transcriptional disruption at *Slc15a2*. Current work being conducted in our lab to define this phenomenon will include Chromatin immuno-precipitation (ChIP) analysis to evaluate histone modifications at the locus. We are also generating a "tunable" construct with drug-inducible promoters at *Slc15a2* in F9 cells. These experiments will further elucidate the precise roles played by transcriptional interference and RNA interference in the disruption of *Slc15a2* expression and function.

A similar disruption of *MCC* expression was recently found to be a result of three intronic novel TE integrants[5]. Our analysis of Complete Genomics whole genome data found that the L1-beta was present in 9% of the samples, the L1-alpha in 21% and the *Alu* element in 2% - indicating that at least one of these TEs is present in over 30% of the samples available. Moreover, TE polymorphisms present at high allele frequencies largely would not be expected to result in a significant gene disruption in essential tumor suppressor genes in the human genome, as was reported[3]. Furthermore, this down-regulation of *MCC* was indicated to be linked to liver cancer in the human population[5]. Because over 30% of the human population does not develop liver cancer as a result of the reported TE-mediated disruption of *MCC*, we were critical of the finding.

Nonetheless, we studied this locus to understand whether the putative TE-mediated disruption actually occurs at *MCC* and to develop a system in which similar mechanism studies can be conducted. We first obtained 10 human lymphoblastoid cell lines and 22 liver cancer samples to confirm the presence or absence of each of the three TEs via a transposon-junction PCR assay. Next, we checked relative expression of the gene to correlate the presence of one of the reported TEs with down-regulation. And we anticipated, in our preliminary qRT-PCR analysis of 10 human lymphoblastoid cell lines and 22 liver cancer samples, variable expression levels of *MCC* were not associated with the presence of one or none of the intronic TEs.

We recognize that in order to confirm our results, more samples are needed to achieve biological and statistical significance. However, this analysis corroborates our hypothesis that TE-mediated down-regulation may not occur at the *MCC* locus in humans. In their initial report of the phenomenon, the Faulkner group may have simply found the few samples in which the presence of the TE could be related to reduction in *MCC* expression in liver cancer samples.

However, our analysis did not confirm their result in studying additional samples. We conclude that TE-mediated disruption largely does not occur at *MCC.*

Previous reports have shown the potency of transposons to result in gene disruption within the genome. Our own studies at both loci in different organisms, *Slc15a2* in mice and *MCC* in humans, reveal the importance of a case-by-case analysis to identify novel instances of TE-mediated gene disruptions. These results provide insights into the conditions in which transposons are capable of initiating transcriptional instability and potentially leading to disease. Known examples of where TE insertions have induced a disease phenotype include the well-documented A^{vy} gene in mice, in which IAP integration induced overexpression of the *A* gene, leading to obesity and type II diabetes[6]. Our own analysis of *Slc15a2* proposes a novel model of transcriptional regulation by antisense transcripts initiated by new TE integrations (Figure 14). This field of research remains largely unexplored and to the best of our knowledge, *Slc15a2* could become one of the well-defined genes at which this mechanism of disruption in gene transcription and function occurs.

Due to the prevalence of transposons within the mammalian genome and potency of various TE integrants to induce genetic instability, a careful analysis must be undertaken to determine how changes in gene expression result from new TE integrations, and then to follow up on well-documented cases to determine the underlying molecular mechanisms.

Acknowledgements

I would like to thank my project advisor, Dr. David Symer for experimental design and support through the entirety of this research project and the completion of my thesis. With his encouragement and advice, I have learned many fundamental skills that have helped me gain confidence as an undergraduate researcher. I would also like to thank Dr. Dandan He, Dr. Jingfeng Li, and Dr. Keiko Akagi for guidance during experimental work. Each was personally invested in teaching me throughout the course of the two years I have worked in the Symer lab.

References

1. Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860-921.

2. Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., *et al.*2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520-562.

3. Akagi, K., Li, J., Symer, D.E. 2013. How do mammalian transposons induce genetic variation? A conceptual framework. *Bioessays* **35:** 397-407.

4. Li, J., Akagi, K., Hu, Y., Trivett, A. L., Hlynialuk, C., Swing, D. A., Volfovsky, N., Morgan, T. C., Golubeva, Y., Stephens, R. M., Smith, D. E., and Symer, D. E. 2012. Mouse endogenous retroviruses can trigger premature transcriptional termination at a distance. *Genome Research* **22:** 870-884.

5. Shukla, R., Upton, K. R., Muñoz-Lopez, M., Gerhardt, D. J., Fisher, M. E., Nyugen, T., Brennan, P. M, Baillie, J. K., Collino, A., Ghisletti, S., Sinha, S. *et al*. 2013. Endogenous retrotransposition activates oncogenic pathways in hepatocellular carcinoma. *Cell* **153:** 101-111.

6. Duhl, D. M., Vrieling, H., Miller, K. A., Wolff, G. L., and Barsh, G. S. 1994. Neomorphic agouti mutations in obese yellow mice. *Nature Genetics* **8:** 59-65.

7. Druker, R., Bruxner, T. J., Lehrbach, N. J., and Whitelaw, E. 2004. Complex patterns of transcription at the insertion site of a retrotransposons in the mouse. *Nucleic Acids Research* **32:** 5800-5808.

8. Howard, G., Eiges, R., Gaudent, F., Jaenisch, R., and Eden, A. 2008. Activation and transposition of endogenous retroviral elements in hypomethylation induced tumors in mice. *Oncogene* **27:** 404-408.

9. Rodić, N., and Burns, K. H. 2013. Long Interspersed Element-1 (LINE-1): Passenger or Driver in Human Neoplasms? *PLOS Genetics* **9:** e1003402.

10. Akagi, K., Li, J., Stephens, R. M., Volfovsky, N., and Symer, D.E. 2008. Extensive variation between mouse strains due to endogenous L1 retrotransposition. *Genome Research* **18:** 869-880.

11. Kazazian Jr., H. H., Goodier, J. L. 2002. LINE Drive: Retrotransposition and Genomic Instability. *Cell* **110:** 277-280.

12. Zhang, Y., Romanish, M. T., Mager, D. L. 2011. Distributions of Transposable Elements Reveal Hazardous Zones in Mammalian Introns. *PLOS Computational Biology* **7:** e1002046.

13. Li, J., Kannan, M., Trivett, A.L, Liao, H., Wu, X., Akagi, K., and Symer, D.E. 2014. An antisense promoter in mouse L1 retrotransposon open reading frame-1 initiates expression of diverse fusion transcripts and limits retrotransposition. *Nucleic Acids Research* **doi:** 10.1093/nar/gku091.

14. Babatz, T. D. and Burns, K. H. 2013. Functional impact of the human mobilome. *Current Opinion in Genes & Dev.* **23:** 264-270.

15. Ostertag, E. M., Goodier, J. L., Zhang, Y., Kazazian Jr., H. H. 2003. SVA elements are nonautonomous retrotransposons that cause disease in humans. *The American Journal of Human Genetics* **73:** 1444-1451.

16. Miki, Y., Nishisho, I., Horri, A., Miyoshi, Y., Utsunomiya, J., Kinzler, K. W., Vogelstein, B., and Nakamura, Y. 1992. Disruption of the *APC* gene by a retrotransposable insertion of L1 sequence in a colon cancer. *Cancer Research* **52:** 643-645.

17. Faulkner, G. J., Kimura, Y., Daub, C.O., Wani, S., Plessy, C., Irvine, K.M, Schroder, K., Cloonan, N., Steptoe, A. L., Lassman, T., *et al.* 2009. The regulated retrotransposons transcriptome of mammalian cells. *Nature Genetics* **41:** 563-571.

18. Lau, C., Sun, T., Ching, A., He, M., Li, J., Wong, A. M., Co, N. N., Chan, A., Li, P., Lung, R., Tong, J., *et al.* 2014. Viral-human chimeric transcript predisposed risk to liver cancer development and progression. *Cancer Cell* **25:** 335-349.

19. Heikenwalder, M., and Protzer, U. 2014. LINE(1)s of evidence in HBV-driven liver cancer. *Cell Host & Microbe* **15:** 249-250.

20. Vasicek, T. J., Zeng, L., Guan, X. J., Zhang, T., Constantini, F., and Tilghman, S. M. 1997. Two dominant mutations in the mouse fused gene are the result of transposon insertions. *Genetics* **147:** 777-786.

21. Morgan, H., Sutherland, H., Martin, D., and Whitelaw, E. 1999. Epigenetic inheritance at the agouti locus in the mouse. *Nature Genetics* **23:** 314-318.

22. Kuehner, J. N., Pearson, E. L., and Moore, C. 2011. Unravelling the means to an end: RNA polymerase II transcription termination. *Nature Rev.: Mol. Cell Bio.* **12:** 283-294.

23. Kadonaga, J. T. 2004. Regulation of RNA polymerase II transcription by sequence-specific DNA binding factors. *Cell* **116:** 247-257.

24. Mata, J., Marguerat, S., and Bähler, J. 2009. Post-transcriptional control of gene expression: a genome-wide perspective. *Trends in Biochem. Sci.* **30:** 506-514.

25. Conley, A. B. and Jordan, I. K. 2012. Epigenetic regulation of human *cis*-natural antisense transcripts. *Nucleic Acids Research* **40:** 1438-1445.

26. Katayama, S. *et al.*2005. Antisense transcription in the mammalian transcriptome. *Science* **309:** 1564.

27. Mattick, J. S. 2004. RNA regulation: a new genetics? *Nature Review Genetics* **5:** 316-323.

28. Morrissy, A. S., Griffith, M. and Marra, M. A. 2011. Extensive relationship between antisense transcription and alternative splicing in the human genome. *Genome Research* **21:** 1203-1212.

29. Kim, D. S. and Hahn, Y. 2010. Human-specific antisense transcripts induced by the insertion of transposable element. *Inter. Jour. of Mol. Medicine* **26:** 151-157.

30. Shearwin, K. E., Callen, B. P., and Egan, J.B. 2005. Transcriptional interference – a crash course. *Trends in Genetics* **21:** 339-345.

31. Conley, A. B., Miller, W. J., and Jordan, I. K. 2008. Human *cis* natural antisense transcripts initiated by transposable elements. *Trends in Genetics* **24:** 53-56.

32. Meister, G. and Tuschl, T. 2004. Mechanisms of gene silencing by double-stranded RNA. *Nature* **431:** 343-349.

33. Yoder, J. A., Walsh, C.P, and Bestor, T.H. 1997. Cytosine methylation and the ecology of intragenomic parasites. *Trends in Genetics* **13:** 335-340.

34. Bestor, T.H. 2000. The DNA methyltransferases of mammals. *Human Molecular Genetics* **9:** 2395-2402.

35. Moore, L. D., Le, T., and Fan, G. 2013. DNA methylation and its basic function. *Neuropsychopharmacology Reviews* **38:** 23-38.

36. Lei, H., Oh, S. P., Okano, M., Jüttermann, R., Goss, K. A., Jaenish, R., and Li, E. 1996. De novo DNA cytosine methyltransferase activities in mouse embryonic stem cells. *Development* **122:** 3195-3205.

37. Smallwood, S. A. and Kelsey, G. 2012. De novo DNA methylation: a germ cell perspective. *Trends in Genetics* **28:** 33-42.

38. Damelin, M. and Bestor, T. H. 2007. Biolgoical functions of DNA Methyltransferase 1 require its methyltransferase activity. *Mol Cell Biol* **27:** 3891-3899.

39. Li, E., Bestor, T. H., and Jaenisch, R. 1992. Targeted mutation of the DNA Methyltransferase gene results in embryonic lethality. *Cell* **69:** 915-926.

40. Tucker, K. L., Beard, C., Dausman, J., Jackson-Grusby, L., Laird, P. W., Lei, H., Li, E., and Jaenisch, R. 1996. Germ-line passage is required for establishment of methylation and expression patterns of imprinted but not of nonimprinted genes. *Genes & Development* **10:** 1008-1020.

41. Gaudet, F., Hodgson, J. G., Eden, A., Jackson-Grusby, L., Dausman, J., Gray, J. W., Leonhardt, H., and Jaenisch, R. 2003. Induction of tumors in mice by genomic hypomethylation. *Science* **300:** 489-492.

42. Smith, D. E., Clémençon, B., and Hediger, M. A. 2013. Proton-coupled oligopeptide transporter family SLC15: physiological, pharmacological and pathological implications. *Molecular Aspects of Medicine* **34:** 323-336.

43. Nellaker, C., Keane, T. M., Yalcin, B., Wong, K., Agam, A., Belgard, T. G., Flint, J., Adams, D. J., Frankel, W. N., and Ponting, C. P. 2012. The genomic landscape shaped by selection on transposable elements across 18 mouse strains. *Genome Biology* **doi:** 10.1186/gb-2012-13-6-r45.

44. Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., Pimentel, H., Salzberg, S. L., Rinn, J. L., Pachter, L. 2012. Differential gene and transcript analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols* **7:** 562-578.

45. Fukuyama, R., Niculaita, R., Ng, K. P., Obusez, E., Sanchez, J., Kalady, M., Aung, P. P., Casey, G., and Sizemore, N. 2008. *Mutated in Colorectal Cancer*, a putative tumor suppressor for serrated colorectal cancer, selectively represses β-catenin-dependent transcription**.** *Oncogene* **27:** 6044-6055.

46. Patterson, K., Molloy, L., Qu, W., and Clark, S. 2011. DNA methylation: bisulfite modification and analysis. *Journal of Visualized Experiments* **56:** 3170.